

我这一周总结了一下地图里面现有的问题。

一、已有工作：

1.数据：

系统使用的数据还是汇总每个小时的数据，许多数据属性是模拟出来的。所以下一步重点是分析真实数据。

2.系统：

大家普遍觉得系统偏展示。从开发的角度，原因如下：

1. 因为先前一直按照大屏的需求出发，所以会优先考虑展示性的工作。下一步就会从数据分析入手。
2. 系统没有交互，而交互是数据分析关键一步。所以这也是接下来的工作重点之一。

二、下一步解决方案：

仔细想了一下，下一步的问题定位：

1. 从大规模的交易中寻找异常交易。（未来重点）
2. 展现特定城市与区域的形象。（完成一部分）
3. 反应交易趋势。（已有工作）

1. 从大规模的交易中寻找异常交易。

现在已经获得了真实交易数据，放在 FTP 目录下：/FTP2/Data/淘宝交易地图数据/
数据格式为：

ali_date	支付宝支付时间
auction_price	商品单价(元)
buy_amount	购买数量
aa_prov	商品所属省份（起始地）
aa_city	商品所属城市
cat_id	叶子类目 ID
cat_name	叶子类目名
cat1	一级类目 ID
name1	一级类目名称
lgo_prov	物流订单 省份（目的地）
lgo_city	物流订单 城市

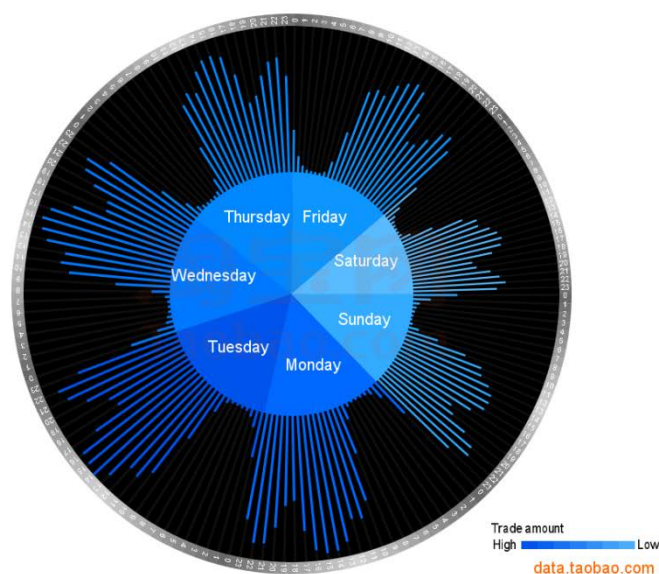
一共有两天的数据：一个是周一的，一个是周日的。每份数据有 20,000,000 条左右（周日略少，1900 万，周一有 2100 万）。数据的维度如上所示，每份数据时间跨度在一天之内。

在与淘宝数据开发人员的交流中，我了解到了如下信息：交易中存在许多虚假交易。这个正好可以作为我们的一个问题来研究。

我们先抽取了金额小于 0.01 元的交易，然后按照时间排序。在 9 月 16 日 3 点左右发现如下交易序列：

E7 福建											
	A	B	C	D	E	F	G	H	I	J	K
6019	2012-9-16 2:56	0.01	1	中山	广东	50003317	网站ID注册	50014811	网店/网络	广东省	广州市
6020	2012-9-16 2:56	0.01	1	重庆	重庆	50003317	网站ID注册	50014811	网店/网络	江苏省	苏州市
6021	2012-9-16 2:57	0.01	1	杭州	浙江	50012708	其它纸品	50018004	电子词典/	四川省	成都市
6022	2012-9-16 2:57	0.01	1	杭州	浙江	50012708	其它纸品	50018004	电子词典/	四川省	成都市
6023	2012-9-16 2:57	0.01	1	杭州	浙江	50012708	其它纸品	50018004	电子词典/	四川省	成都市
6024	2012-9-16 2:57	0.01	1	杭州	浙江	50012708	其它纸品	50018004	电子词典/	四川省	成都市
6025	2012-9-16 2:57	0.01	1	青岛	山东	50007782	其他游戏币	50011665	网游装备/	\N	\N
6026	2012-9-16 2:57	0.01	1	杭州	浙江	50012708	其它纸品	50018004	电子词典/	四川省	成都市
6027	2012-9-16 2:57	0.01	1	杭州	浙江	50012708	其它纸品	50018004	电子词典/	四川省	成都市
6028	2012-9-16 2:57	0.01	1	杭州	浙江	50012708	其它纸品	50018004	电子词典/	四川省	成都市
6029	2012-9-16 2:57	0.01	1	杭州	浙江	50012708	其它纸品	50018004	电子词典/	四川省	成都市
6030	2012-9-16 2:57	0.01	1	杭州	浙江	50012708	其它纸品	50018004	电子词典/	四川省	成都市
6031	2012-9-16 2:59	0.01	1	杭州	浙江	50012708	其它纸品	50018004	电子词典/	四川省	成都市
6032	2012-9-16 2:59	0.01	1	杭州	浙江	50012708	其它纸品	50018004	电子词典/	四川省	成都市
6033	2012-9-16 2:59	0.01	1	杭州	浙江	50012708	其它纸品	50018004	电子词典/	四川省	成都市
6034	2012-9-16 2:59	0.01	1	杭州	浙江	50012708	其它纸品	50018004	电子词典/	四川省	成都市
6035	2012-9-16 2:59	0.01	1	广州	广东	50025955	其他	50016348	清洁/卫浴	山东省	烟台市
6036	2012-9-16 2:59	0.01	1	杭州	浙江	50012708	其它纸品	50018004	电子词典/	四川省	成都市
6037	2012-9-16 2:59	0.01	1	杭州	浙江	50012708	其它纸品	50018004	电子词典/	四川省	成都市
6038	2012-9-16 2:59	0.01	1	杭州	浙江	50012708	其它纸品	50018004	电子词典/	四川省	成都市
6039	2012-9-16 2:59	0.01	1	杭州	浙江	50012708	其它纸品	50018004	电子词典/	四川省	成都市
6040	2012-9-16 2:59	0.01	1	杭州	浙江	50012708	其它纸品	50018004	电子词典/	四川省	成都市
6041	2012-9-16 3:00	0.01	1	杭州	浙江	50012708	其它纸品	50018004	电子词典/	四川省	成都市
6042	2012-9-16 3:00	0.01	1	杭州	浙江	50012708	其它纸品	50018004	电子词典/	四川省	成都市
6043	2012-9-16 3:00	0.01	1	杭州	浙江	50012708	其它纸品	50018004	电子词典/	四川省	成都市
6044	2012-9-16 3:00	0.01	1	杭州	浙江	50012708	其它纸品	50018004	电子词典/	四川省	成都市
6045	2012-9-16 3:00	0.01	1	杭州	浙江	50012708	其它纸品	50018004	电子词典/	四川省	成都市
6046	2012-9-16 3:00	0.01	1	莆田	福建	50007716	T-天龙八音	50011665	网游装备/	\N	\N
6047	2012-9-16 3:00	0.01	1	杭州	浙江	50012708	其它纸品	50018004	电子词典/	四川省	成都市
6048	2012-9-16 3:00	0.01	1	杭州	浙江	50012708	其它纸品	50018004	电子词典/	四川省	成都市
6049	2012-9-16 3:00	0.01	1	杭州	浙江	50012708	其它纸品	50018004	电子词典/	四川省	成都市
6050	2012-9-16 3:01	0.01	1	杭州	浙江	50012708	其它纸品	50018004	电子词典/	四川省	成都市
6051	2012-9-16 3:01	0.01	1	武汉	湖北	50020126	网络店铺什	50025111	本地化生产	广东省	深圳市
6052	2012-9-16 3:01	0.01	1	杭州	浙江	50012708	其它纸品	50018004	电子词典/	四川省	成都市
6053	2012-9-16 3:01	0.01	1	杭州	浙江	50012708	其它纸品	50018004	电子词典/	四川省	成都市
6054	2012-9-16 3:01	0.01	1	杭州	浙江	50012708	其它纸品	50018004	电子词典/	四川省	成都市
6055	2012-9-16 3:01	0.01	1	杭州	浙江	50012708	其它纸品	50018004	电子词典/	四川省	成都市
6056	2012-9-16 3:01	0.01	1	杭州	浙江	50012708	其它纸品	50018004	电子词典/	四川省	成都市
6057	2012-9-16 3:01	0.01	1	杭州	浙江	50012708	其它纸品	50018004	电子词典/	四川省	成都市
6058	2012-9-16 3:01	0.01	1	杭州	浙江	50012708	其它纸品	50018004	电子词典/	四川省	成都市
6059	2012-9-16 3:01	0.01	1	杭州	浙江	50012708	其它纸品	50018004	电子词典/	四川省	成都市
6060	2012-9-16 3:02	0.01	1	杭州	浙江	50012708	其它纸品	50018004	电子词典/	四川省	成都市

然而，从下图可以看出周一到周日交易最少的时间段集中在 2:00a.m-6:00a.m。而上面的表格显示凌晨 3 点的时间段出现了大量（150 笔以上）、小额（1 分钱以下）、单笔只包含一件商品的交易。而且买的东西仅仅是纸品。



虽然数据不包含用户 ID，但是所有交易的地点相同：从杭州到成都。如此密集的交易使得

有充分理由怀疑杭州的是同一个卖家。推测可能是卖家在刷信誉。

另外，在使用其他工具整理数据的时候，也发现了另外一些购买模式：
其中第一到七列分别表示：交易时间，单笔金额，单笔数目，卖家地点，叶子类目 ID，类目 ID，买家地点。

2012-09-24 04:00:15	0.01	1	金华	50003317	50014811	沈阳市
2012-09-24 04:00:21	0.01	1	宜昌	50023807	50025111	阿里地区
2012-09-24 04:00:32	0.01	1	温州	50003317	50014811	红河哈尼族彝族自治州
2012-09-24 04:00:32	0.01	1	惠州	50003317	50014811	红河哈尼族彝族自治州
2012-09-24 04:00:32	0.01	1	无锡	50003317	50014811	红河哈尼族彝族自治州
2012-09-24 04:00:32	0.01	1	武汉	50003317	50014811	红河哈尼族彝族自治州
2012-09-24 04:00:32	0.01	1	东莞	50009047	50010404	红河哈尼族彝族自治州
2012-09-24 04:00:32	0.01	1	金华	50003317	50014811	红河哈尼族彝族自治州
2012-09-24 04:00:32	0.01	1	东莞	50009047	50010404	红河哈尼族彝族自治州
2012-09-24 04:00:32	0.01	1	东莞	50009047	50010404	红河哈尼族彝族自治州
2012-09-24 04:00:32	0.01	1	东莞	50009047	50010404	红河哈尼族彝族自治州
2012-09-24 04:00:32	0.01	1	东莞	50009047	50010404	红河哈尼族彝族自治州
2012-09-24 04:00:32	0.01	1	东莞	50009047	50010404	红河哈尼族彝族自治州
2012-09-24 04:00:41	0.01	1	宜昌	50023807	50025111	固原市
2012-09-24 04:00:47	0.01	1	东莞	50009047	50010404	重庆市

可以看出在 9 月 24 日凌晨 4:00:32 的时刻，突然从红河哈尼族彝族自治州出现了 10 条交易。引起我注意的是买家所在地都是同一地点，而卖家则各不相同。因为没有用户 ID，所以我不能判断是否是同一个买家。但是有趣的是买家的地点是一个小城市，该地的交易量在全部交易的比重并不高，而瞬间出现如此密集的交易显然是值得怀疑的。初步判断可能是某一个买家在帮人刷信誉，而且我觉得他可能并不是位于“红河哈尼族彝族自治州”，他只是使用了一个假的地址而已。

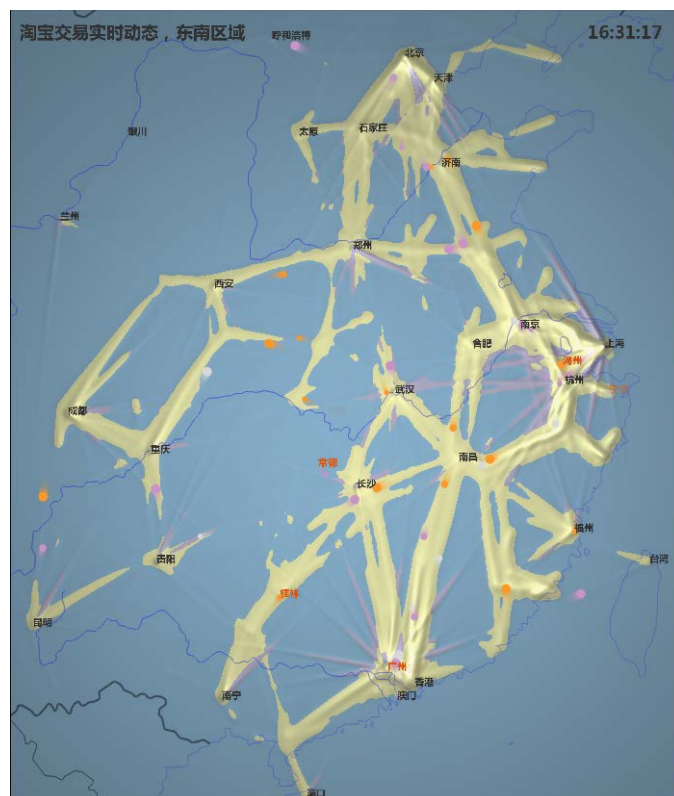
另外，除了刷信誉的交易以外。肯定还可以发现其他特定交易方式。但是我们现在仅仅经过初步的探索还不知道，需要进一步发掘。

因此，我对这类数据的想法是：

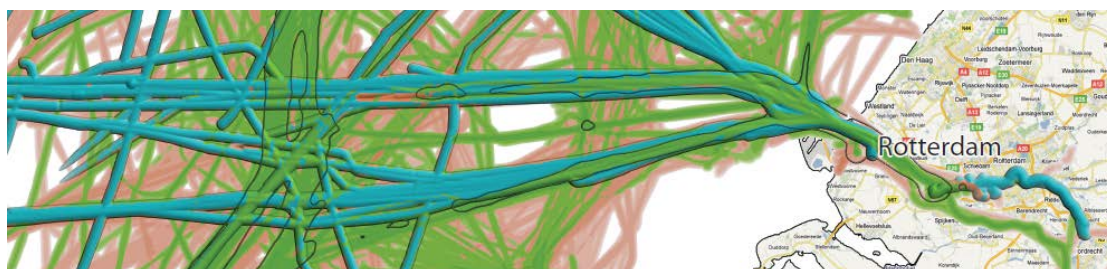
1. 对数据与处理，去除掉冗余和无用信息。（数据中有很多信息不太全的，或许应该剔除掉）
2. 先采用某种聚类方式尝试处理数据。如果两千万条数据太大找不到合适方法聚类，就先选取一个小时的数据聚类。平均下来一个小时有 100 万条左右。
3. 以上数据可能和频率是有关系的，所以是不是可以联系到短时傅里叶分解来分析。
4. 虽然以上数据维度只有 7，但是包含了许多隐藏维度。比如地点是不是一级城市，交易频率也是不能直接从这几个属性中得出的，需要联系其他的交易才行。是否可以通过某种概率模型来分析？

2. 展现特定城市与区域的形象。

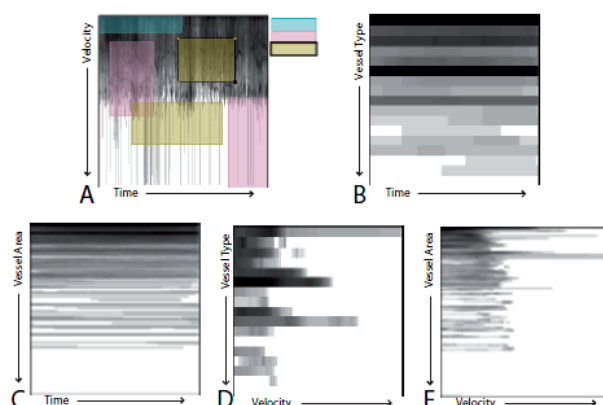
在已有的系统中我开发了花瓣图：



可以继续参考这篇文章：[Composite Density Maps for Multivariate Trajectories TVCG 2011](#)



其数据是真实的轨迹数据。作者使用了类似传输函数调整的界面提供给用户框选所要显示的信息。比如：选择特定速度，特定类型的船只。



最后作者还利用这些可视化形式进行了船只风险分析等等。
我们的系统之所以被认为是偏展示，很大一部分原因是未添加用户交互的部分，比如上图所示的交互选取的模块。